IAndMV

Inductive Anomaly Detection with Missing Value

E14051350 吳岱桀

01

Initial Framework

Anomaly Detection

Introduction

Binary or Multi-class classification:

Classify the data as normal vs anomaly or normal vs different classes of anomaly

Imbalanced Problem:

Anomaly data is rare in the real life

e.g. Medical, Fraud

Core Method

Tabular to Graph

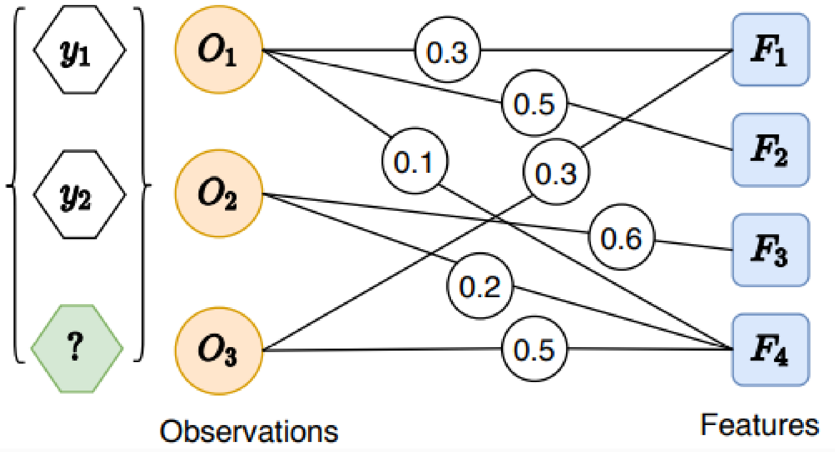
Data Matrix
with Missing Values

	F_1	F_2	F_3	F_4
O_1	0.3	0.5	NA	0.1
O_2	NA	NA	0.6	0.2
O_3	0.3	NA	NA	0.5

Labels

Y
y_1
y_2
?

$$\textsc{Init}(v) = \begin{cases} 1 & v \in \mathcal{V}_D \\ \textsc{OneHot} & v \in \mathcal{V}_F \end{cases}$$



"Handling Missing Data with Graph Representation Learning"

Core Method

Tabular to Graph

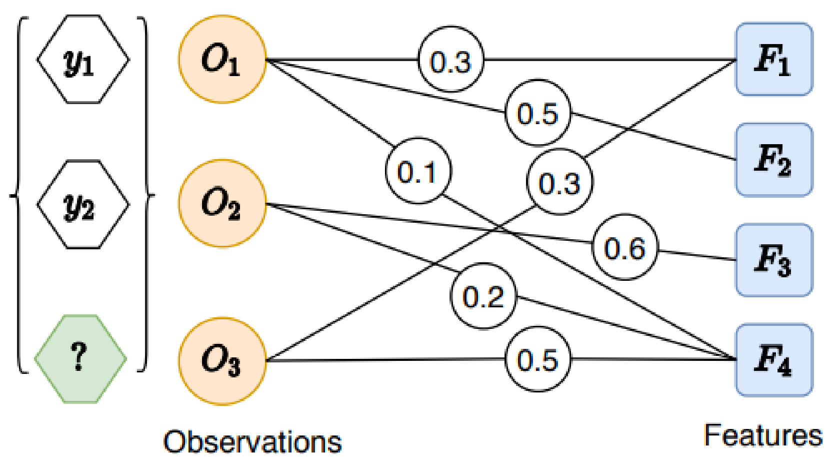
Used as an **Imputation Method**.

It shows that predict the label with embedding directly is better than predict after imputation

Embedding

$$\mathbf{n}_v^{(l)} = \text{AGG}_l\Big(\sigma(\mathbf{P}^{(l)} \cdot \text{CONCAT}(\mathbf{h}_v^{(l-1)}, \mathbf{e}_{uv}^{(l-1)}) \mid \forall u \in \mathcal{N}(v, \mathcal{E}_{drop}))\Big)$$

GAT



$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)}$$

$$\vec{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right)$$

$$\mathbf{h}_v^{(l)} = \sigma(\mathbf{Q}^{(l)} \cdot \text{CONCAT}(\mathbf{h}_v^{(l-1)}, \mathbf{n}_v^{(l)}))$$

$$\mathbf{e}_{uv}^{(l)} = \sigma(\mathbf{W}^{(l)} \cdot \text{CONCAT}(\mathbf{e}_{uv}^{(l-1)}, \mathbf{h}_u^{(l)}, \mathbf{h}_v^{(l)}))$$

How to detect?

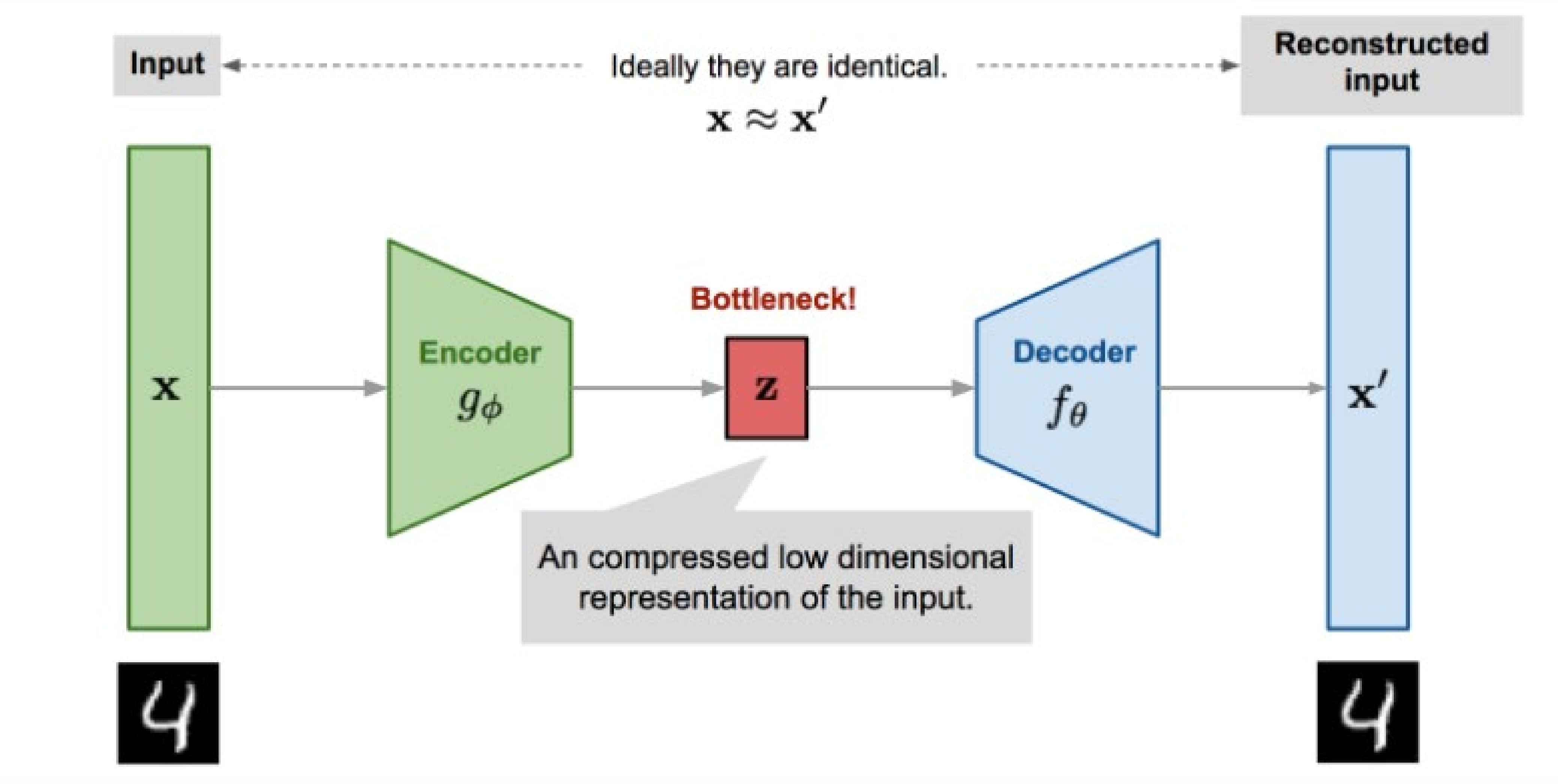
(1) Node Classification

Use node embedding to predict whether it’s anomaly or not

(2) Reconstruction Error

Use embedding to reconstruct adjacency matrix and node feature

Reconstruction Error



Anomaly Detection

(1) Node Classification

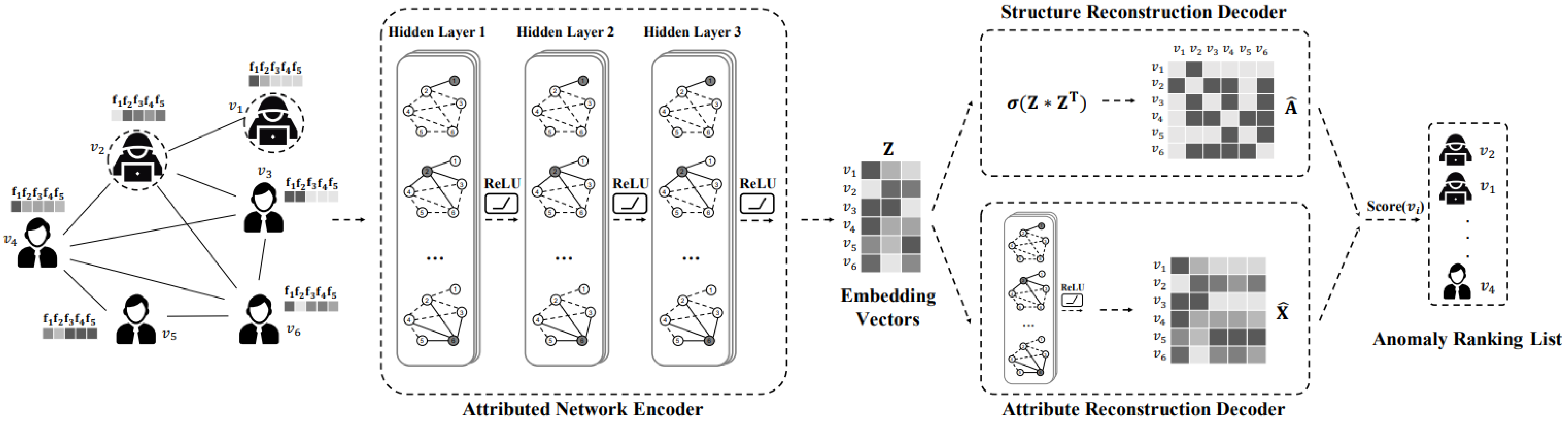
$$\hat{\mathbf{Y}}_u = \mathbf{O}_{node}(\hat{\mathbf{D}}_{u\cdot})$$

(2) Reconstruction Error → Link Prediction

$$\mathbf{O}_{edge}(\textsc{Concat}(\mathbf{h}_u^{(L)}, \mathbf{h}_v^{(L)})) \rightarrow \widehat{A}$$

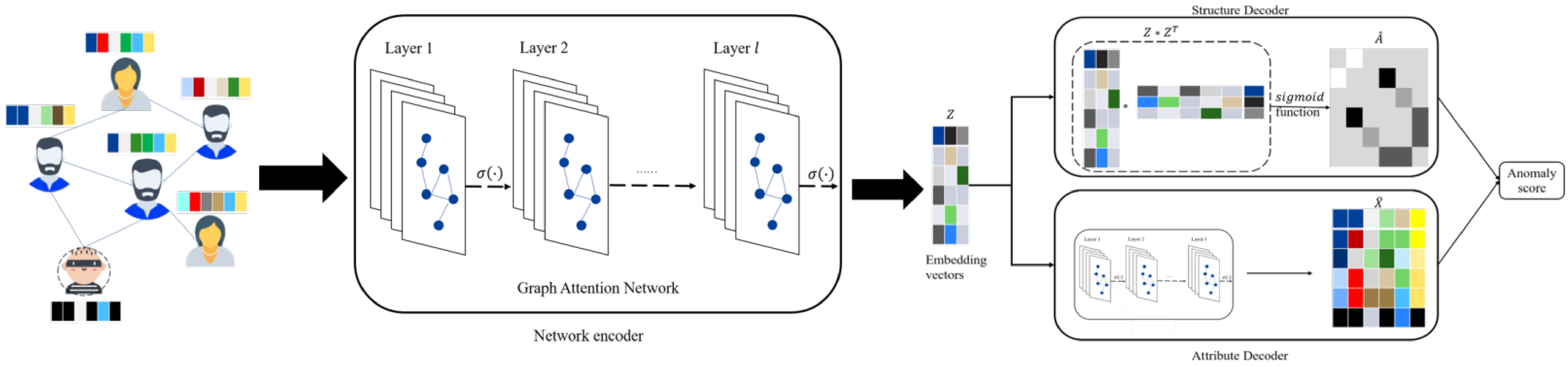
$$\text{Error} = \left\| A - \widehat{A} \right\|_F$$

Previous Research



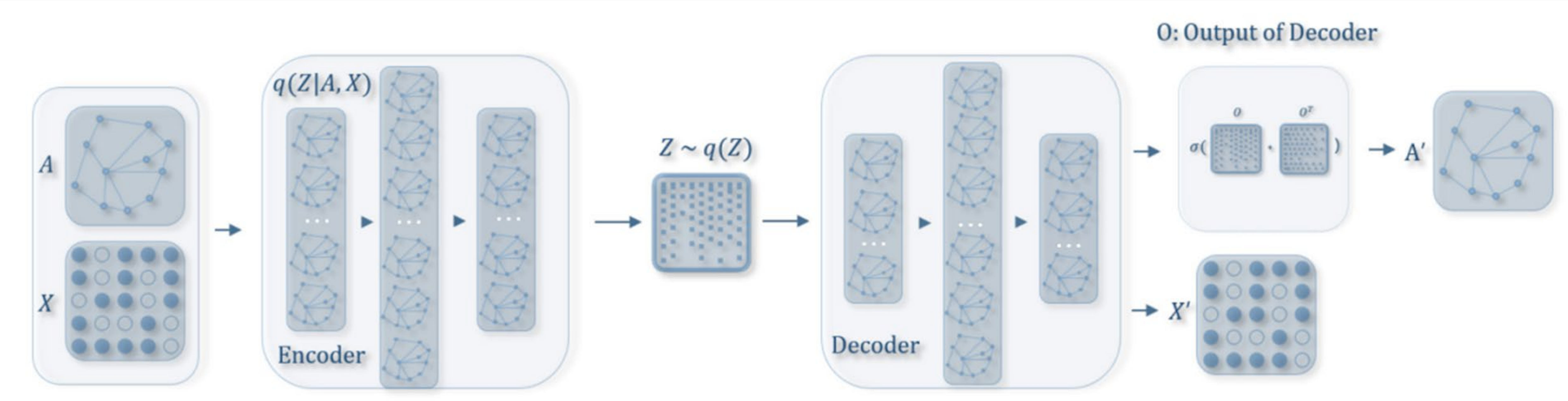
“Deep Anomaly Detection on Attributed Networks”

Previous Research



"GATAE: Graph Attention-based Anomaly Detection on Attributed Networks"

Previous Research



"Learning Graph Embedding With Adversarial Training Methods"

Result(F1 Score)

5 fold test result without missing value

Method	Dataset			
	Arrhythmia	Thyroid	KDD	KDDRev
OC-SVM	45.8	38.9	79.5	83.2
LOF	50	52.7	83.8	81.6
IForest	51.4	63	90.7	87.7
Classification(Ours)	51.7	69.7	92.4	93.5
Reconstruction Error(Ours)	51.3	65	91.3	93

02

Work in Progress

Improvement

Better Embedding

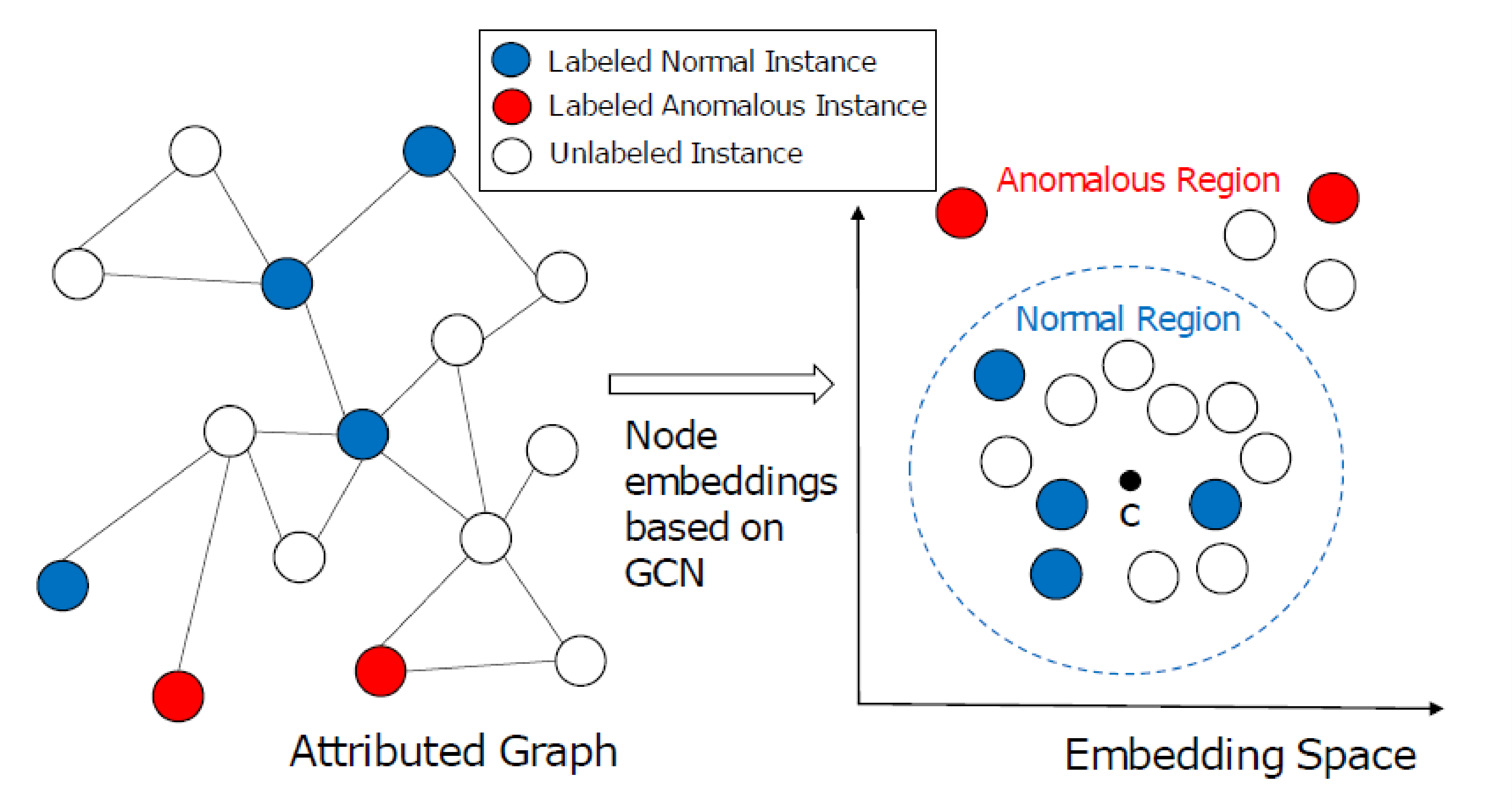
(1) One Class/ Semi-supervised ?

Learn better embedding that can make classifier have clear decision boundary

(2) Can GAN helps?

Helps the to learn better embedding using prior distribution

Semi-supervised



Anomaly Score:

$$a(v_n) := \|\mathbf{h}_n - \mathbf{c}\|^2$$

"Semi-supervised Anomaly Detection on Attributed Graphs"

To Do List & Future Work

Hypersphere Loss(using normal only)

$$\mathcal{L}_{\text{nor}}(\theta) := \frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} a(v_n) = \frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} \|\mathbf{h}_n - \mathbf{c}\|^2$$

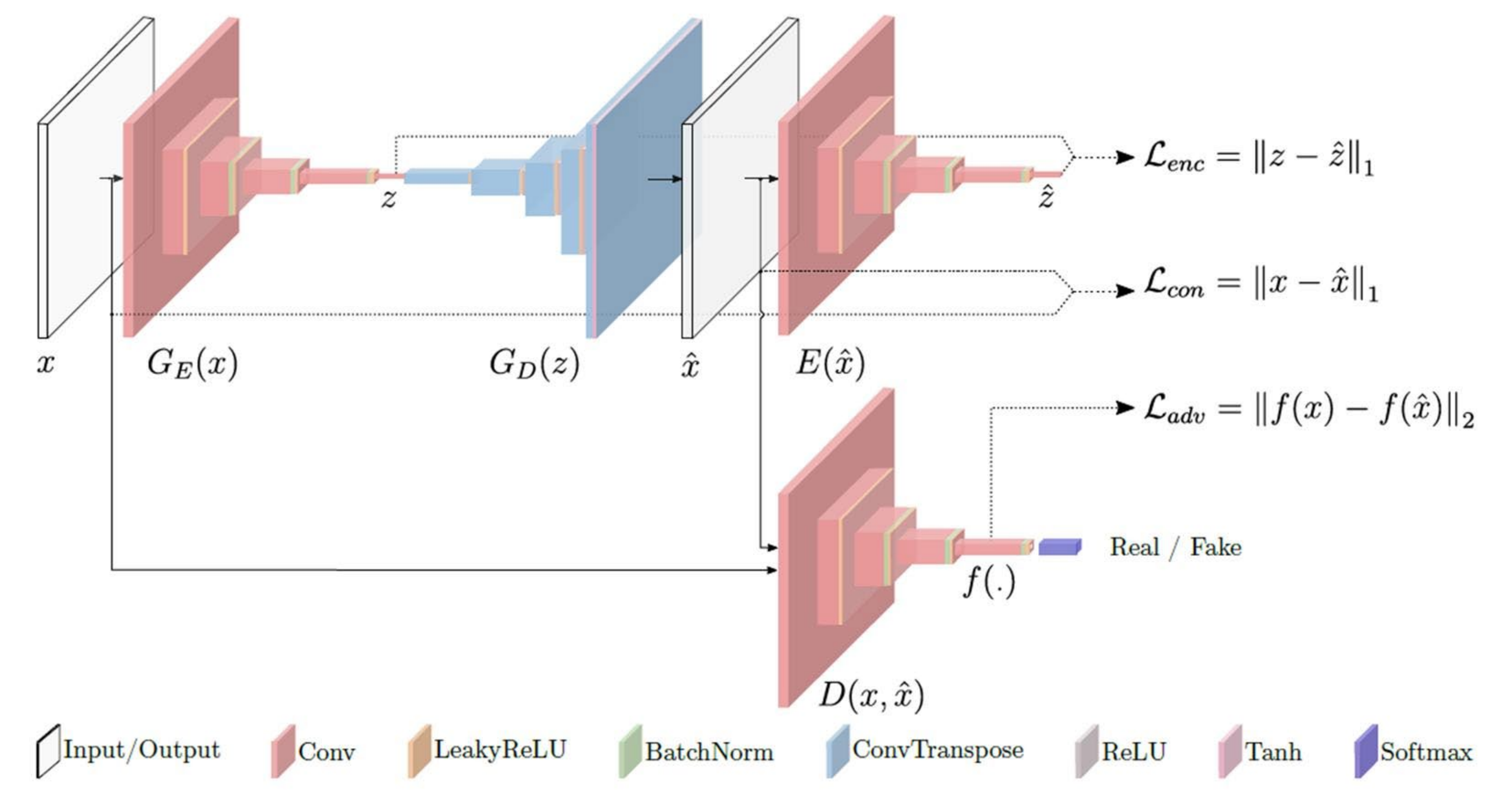
Differential Approximation of AUC

$$\mathcal{R}_{\text{AUC}}(\theta) := \frac{1}{|\mathcal{A}||\mathcal{N}|}\sum_{n\in\mathcal{A}}\sum_{m\in\mathcal{N}} f(a(v_n) - a(v_m))$$

Total Loss

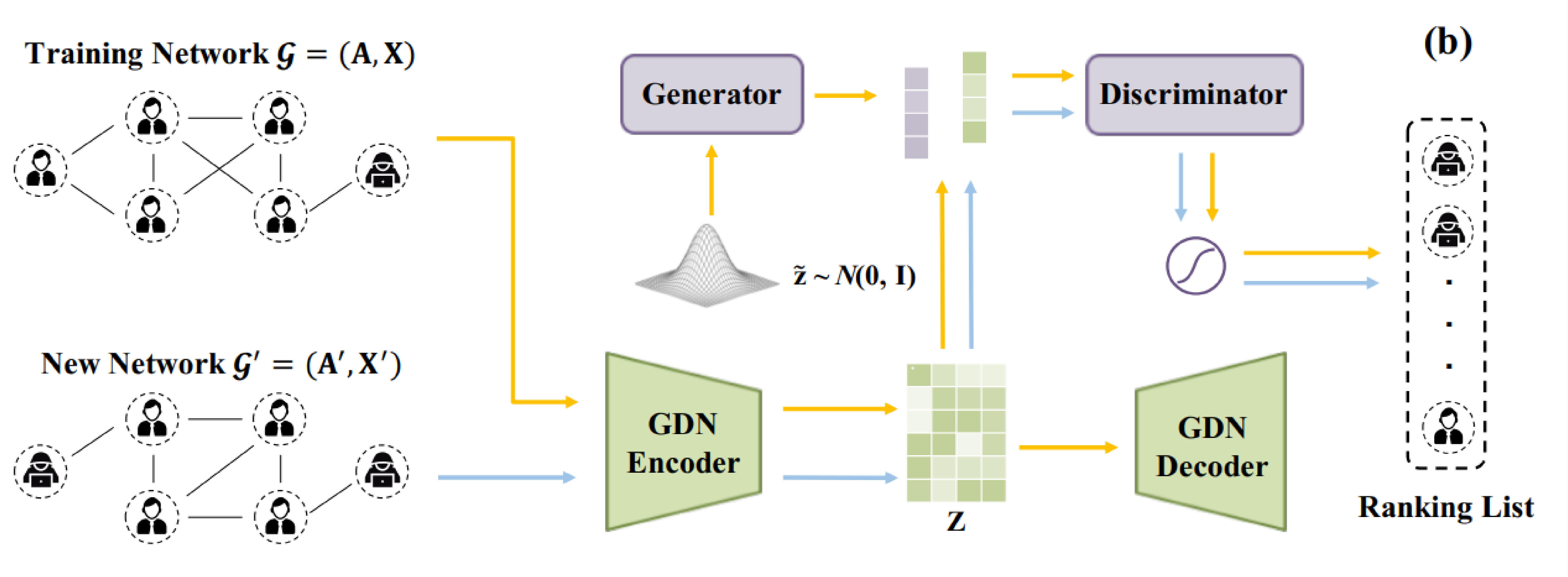
$$\mathcal{L}(\theta) := \mathcal{L}_{\text{nor}}(\theta) - \lambda\mathcal{R}_{\text{AUC}}(\theta)$$

GANomaly



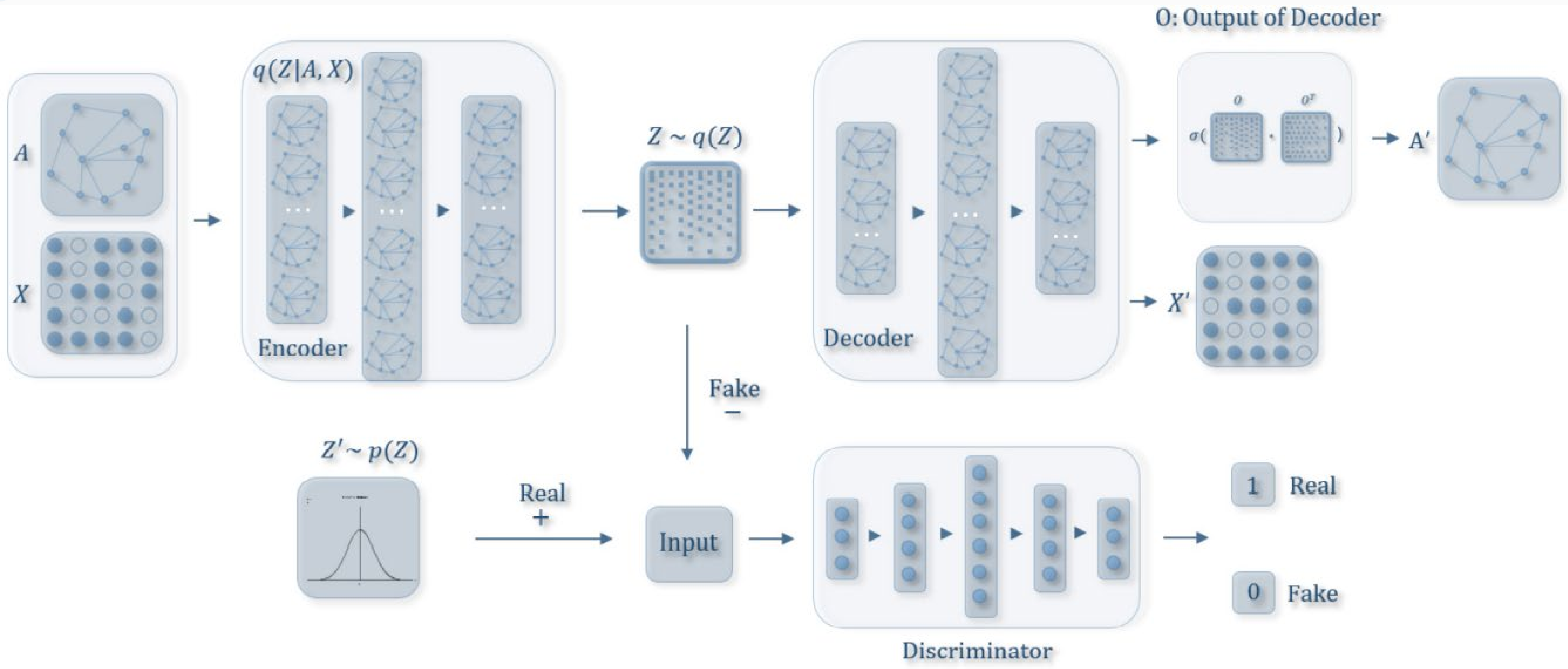
"GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training"

GAN for graph



"Inductive Anomaly Detection on Attributed Networks"

GAN for graph



"Learning Graph Embedding With Adversarial Training Methods"

To Do List
&
Future Work

- Semi-supervised + GAN
- Different Combination of Framework
- Ablation & Missing Ratio Study
- Few-shot & Meta-learning

THANK YOU

E14051350 吳岱桀